arXiv:1103.5537v2 [stat.ML] 30 Mar 2011

Least-Squares Independence Regression for Non-Linear Causal Inference under Non-Gaussian Noise

Makoto Yamada[1], Masahi Sugiyama[1,2], and Jun Sese[3]
[1]Tokyo Institute of Technology, Japan
[2]PRESTO, Japan Science and Technology Agency, Tokyo, Japan
[3]Department of Information Science, Ochanomizu University, Japan
{yamada@sg. sugi@}cs.titech.ac.jp, sesejun@is.ocha.ac.jp

January 18, 2013

Abstract

The discovery of non-linear causal relationship under additive non-Gaussian noise models has attracted considerable attention recently because of their high flexibility. In this paper, we propose a novel causal inference algorithm called *least-squares independence regression* (LSIR). LSIR learns the additive noise model through the minimization of an estimator of the *squared-loss mutual information* between inputs and residuals. A notable advantage of LSIR over existing approaches is that tuning parameters such as the kernel width and the regularization parameter can be naturally optimized by cross-validation, allowing us to avoid overfitting in a data-dependent fashion. Through experiments with real-world datasets, we show that LSIR compares favorably with a state-of-the-art causal inference method.

1 Introduction

Learning *causality* from data is one of the important challenges in the artificial intelligence, statistics, and machine learning communities (Pearl, 2000). A traditional method of learning causal relationship from observational data is based on the linear-dependence Gaussian-noise model (Geiger and Heckerman, 1994). However, the linear-Gaussian assumption is too restrictive and may not be fulfilled in practice. Recently, non-Gaussianity and non-linearity have been shown to be beneficial in causal inference, allowing one to break symmetry between observed variables (Shimizu *et al.*, 2006; Hoyer *et al.*, 2009). Since then, much attention has been paid to the discovery of non-linear causal relationship through non-Gaussian noise models (Mooij *et al.*, 2009).

In the framework of non-linear non-Gaussian causal inference, the relation between a cause X and an effect Y is assumed to be described by $Y = f(X) + E$, where f is a non-linear function and E is non-Gaussian additive noise which is independent of the cause X. Given two random variables X and X', the causal direction between X and X' is decided based on a hypothesis test of whether the model $X' = f(X) + E$ or the alternative model $X = f'(X') + E'$ fits the data well—here, the goodness of fit is measured by independence between inputs and residuals (i.e., estimated noise). Hoyer *et al.* (2009) proposed to learn the functions f and f' by the *Gaussian process* (GP) (Bishop, 2006), and evaluate the independence between the inputs and the residuals by the *Hilbert-Schmidt independence criterion* (HSIC) (Gretton *et al.*, 2005).

However, since standard regression methods such as GP are designed to handle Gaussian noise, they may not be suited for discovering causality in the non-Gaussian additive noise formulation. To cope with this problem,

a novel regression method called *HSIC regression* (HSICR) has been introduced recently (Mooij *et al.*, 2009). HSICR learns a function so that the dependence between inputs and residuals is directly minimized based on HSIC. Since HSICR does not impose any parametric assumption on the distribution of additive noise, it is suited for non-linear non-Gaussian causal inference. Indeed, HSICR was shown to outperform the GP-based method in experiments (Mooij *et al.*, 2009).

However, HSICR still has limitations for its practical use. The first weakness of HSICR is that the kernel width of HSIC needs to be determined manually. Since the choice of the kernel width heavily affects the sensitivity of the independence measure (Fukumizu *et al.*, 2009), lack of systematic model selection strategies is critical in causal inference. Setting the kernel width to the median distance between sample points is a popular heuristic in kernel methods (Schölkopf and Smola, 2002), but this does not always perform well in practice. Another limitation of HSICR is that the kernel width of the regression model is fixed to the same value as HSIC. This crucially limits the flexibility of function approximation in HSICR.

To overcome the above weaknesses, we propose an alternative regression method called *least-squares independence regression* (LSIR). As HSICR, LSIR also learns a function so that the dependence between inputs and residuals is directly minimized. However, a difference is that, instead of HSIC, LSIR adopts an independence criterion called *least-squares mutual information* (LSMI) (Suzuki *et al.*, 2009), which is a consistent estimator of the *squared-loss mutual information* (SMI) with the optimal convergence rate. An advantage of LSIR over HSICR is that tuning parameters such as the kernel width and the regularization parameter can be naturally optimized through cross-validation (CV) with respect to the LSMI criterion.

Furthermore, we propose to determine the kernel width of the regression model based on CV with respect to SMI itself. Thus, the kernel width of the regression model is determined independent of that in the independence measure. This allows LSIR to have higher flexibility in non-linear causal inference than HSICR. Through experiments with real-world datasets, we demonstrate the superiority of LSIR.

2 Dependence Minimizing Regression by LSIR

In this section, we formulate the problem of dependence minimizing regression and propose a novel regression method, *least-squares independence regression* (LSIR).

2.1 Problem Formulation

Suppose random variables $X \in \mathbb{R}$ and $Y \in \mathbb{R}$ are connected by the following additive noise model (Hoyer *et al.*, 2009):

$$Y = f(X) + E,$$

where $f : \mathbb{R} \to \mathbb{R}$ is some non-linear function and $E \in \mathbb{R}$ is a zero-mean random variable independent of X. The goal of dependence minimizing regression is, from i.i.d. paired samples $\{(x_i, y_i)\}_{i=1}^n$, to obtain a function $\widehat{f}$ such that input X and estimated additive noise $\widehat{E} = Y - \widehat{f}(X)$ are independent.

Let us employ a linear model for dependence minimizing regression:

$$f_{\boldsymbol{\beta}}(x) = \sum_{l=1}^{m} \beta_l \psi_l(x) = \boldsymbol{\beta}^\top \boldsymbol{\psi}(x), \tag{1}$$

where m is the number of basis functions, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)^\top$ are regression parameters, $^\top$ denotes the transpose,

and $\boldsymbol{\psi}(x) = (\psi_1(x), \ldots, \psi_m(x))^\top$ are basis functions. We use the Gaussian basis function in our experiments:

$$\psi_l(x) = \exp\left(-\frac{(x - c_l)^2}{2\tau^2}\right),$$

where c_l is the Gaussian center chosen randomly from $\{x_i\}_{i=1}^n$ without overlap and τ is the kernel width.

In dependence minimization regression, we learn the regression parameter $\boldsymbol{\beta}$ as

$$\min_{\boldsymbol{\beta}} \left[I(X, \widehat{E}) + \frac{\gamma}{2}\boldsymbol{\beta}^\top \boldsymbol{\beta} \right],$$

where $I(X, \widehat{E})$ is some measure of independence between X and $\widehat{E}$, and $\gamma \geq 0$ is the regularization parameter for avoiding overfitting.

In this paper, we use the *squared-loss mutual information* (SMI) (Suzuki *et al.*, 2009) as our independence measure:

$$\mathrm{SMI}(X, \widehat{E}) = \frac{1}{2} \iint \left(\frac{p(x, \widehat{e})}{p(x)p(\widehat{e})} - 1 \right)^2 p(x)p(\widehat{e}) \mathrm{d}x \mathrm{d}\widehat{e}.$$

$\mathrm{SMI}(X, \widehat{E})$ is the *Pearson divergence* (Pearson, 1900) from $p(x, \widehat{e})$ to $p(x)p(\widehat{e})$, and it vanishes if and only if $p(x, \widehat{e})$ agrees with $p(x)p(\widehat{e})$, i.e., X and $\widehat{E}$ are independent. Note that ordinary *mutual information* (MI) (Cover and Thomas, 2006),

$$\mathrm{MI}(X, \widehat{E}) = \iint p(x, \widehat{e}) \log \frac{p(x, \widehat{e})}{p(x)p(\widehat{e})} \mathrm{d}x \mathrm{d}\widehat{e}, \tag{2}$$

corresponds to the *Kullback-Leibler divergence* (Kullback and Leibler, 1951) from $p(x, \widehat{e})$ and $p(x)p(\widehat{e})$, and it can also be used as an independence measure. Nevertheless, we adhere to using SMI since it allows us to obtain an analytic-form estimator, as explained below.

2.2 Estimation of Squared-Loss Mutual Information

SMI cannot be directly computed since it contains unknown densities $p(x, \widehat{e})$, $p(x)$, and $p(\widehat{e})$. Here, we briefly review an SMI estimator called *least-squares mutual information* (LSMI) (Suzuki *et al.*, 2009).

Since density estimation is known to be a hard problem (Vapnik, 1998), avoiding density estimation is critical for obtaining better SMI approximators (Kraskov *et al.*, 2004). A key idea of LSMI is to directly estimate the *density ratio*:

$$r(x, \widehat{e}) = \frac{p(x, \widehat{e})}{p(x)p(\widehat{e})},$$

without going through density estimation of $p(x, \widehat{e})$, $p(x)$, and $p(\widehat{e})$.

In LSMI, the density ratio function $r(x, \widehat{e})$ is directly modeled by the following linear model:

$$r_{\boldsymbol{\alpha}}(x, \widehat{e}) = \sum_{l=1}^{b} \alpha_l \varphi_l(x, \widehat{e}) = \boldsymbol{\alpha}^\top \boldsymbol{\varphi}(x, \widehat{e}), \tag{3}$$

where b is the number of basis functions, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_b)^\top$ are parameters, and $\boldsymbol{\varphi}(x, \widehat{e}) = (\varphi_1(x, \widehat{e}), \ldots, \varphi_b(x, \widehat{e}))^\top$ are basis functions. We use the Gaussian basis function:

$$\varphi_l(x, \widehat{e}) = \exp\left(-\frac{(x - u_l)^2 + (\widehat{e} - \widehat{v}_l)^2}{2\sigma^2}\right),$$

where $(u_l, \widehat{v}_l)$ is the Gaussian center chosen randomly from $\{(x_i, \widehat{e}_i)\}_{i=1}^n$ without replacement, and σ is the kernel width.

The parameter $\boldsymbol{\alpha}$ in the model $r_{\boldsymbol{\alpha}}(x, \widehat{e})$ is learned so that the following squared error $J_0(\boldsymbol{\alpha})$ is minimized:

$$\begin{aligned} J_0(\boldsymbol{\alpha}) &= \frac{1}{2} \iint (r_{\boldsymbol{\alpha}}(x, \widehat{e}) - r(x, \widehat{e}))^2 p(x) p(\widehat{e}) \mathrm{d}x \mathrm{d}\widehat{e} \\ &= \frac{1}{2} \iint r_{\boldsymbol{\alpha}}(x, \widehat{e}) p(x) p(\widehat{e}) \mathrm{d}x \mathrm{d}\widehat{e} - \iint r_{\boldsymbol{\alpha}}(x, \widehat{e}) p(x, \widehat{e}) \mathrm{d}x \mathrm{d}\widehat{e} + C, \end{aligned}$$

where C is a constant independent of $\boldsymbol{\alpha}$ and therefore can be safely ignored. Let us denote the first two terms by $J(\boldsymbol{\alpha})$:

$$J(\boldsymbol{\alpha}) = J_0(\boldsymbol{\alpha}) - C = \frac{1}{2} \boldsymbol{\alpha}^\top \boldsymbol{H} \boldsymbol{\alpha} - \boldsymbol{h}^\top \boldsymbol{\alpha}, \tag{4}$$

where

$$\begin{aligned} \boldsymbol{H} &= \iint \boldsymbol{\varphi}(x, \widehat{e}) \boldsymbol{\varphi}(x, \widehat{e})^\top p(x) p(\widehat{e}) \mathrm{d}x \mathrm{d}\widehat{e}, \\ \boldsymbol{h} &= \iint \boldsymbol{\varphi}(x, \widehat{e}) p(x, \widehat{e}) \mathrm{d}x \mathrm{d}\widehat{e}. \end{aligned}$$

Approximating the expectations in $\boldsymbol{H}$ and $\boldsymbol{h}$ by empirical averages, we obtain the following optimization problem:

$$\tilde{\boldsymbol{\alpha}} = \operatorname*{argmin}_{\boldsymbol{\alpha}} \Big[\frac{1}{2} \boldsymbol{\alpha}^\top \widehat{\boldsymbol{H}} \boldsymbol{\alpha} - \widehat{\boldsymbol{h}}^\top \boldsymbol{\alpha} + \lambda \boldsymbol{\alpha}^\top \boldsymbol{\alpha} \Big],$$

where a regularization term $\lambda \boldsymbol{\alpha}^\top \boldsymbol{\alpha}$ is included for avoiding overfitting, and

$$\begin{aligned} \widehat{\boldsymbol{H}} &= \frac{1}{n^2} \sum_{i,j=1}^n \boldsymbol{\varphi}(x_i, \widehat{e}_j) \boldsymbol{\varphi}(x_i, \widehat{e}_j)^\top, \\ \widehat{\boldsymbol{h}} &= \frac{1}{n} \sum_{i=1}^n \boldsymbol{\varphi}(x_i, \widehat{e}_i). \end{aligned}$$

Differentiating the above objective function with respect to $\boldsymbol{\alpha}$ and equating it to zero, we can obtain an analytic-form solution:

$$\widehat{\boldsymbol{\alpha}} = (\widehat{\boldsymbol{H}} + \lambda \boldsymbol{I}_b)^{-1} \widehat{\boldsymbol{h}}, \tag{5}$$

where $\boldsymbol{I}_b$ denotes the b-dimensional identity matrix. It was shown that LSMI is consistent under mild assumptions and it achieves the optimal convergence rate (Kanamori *et al.*, 2009).

Given a density ratio estimator $\widehat{r} = r_{\widehat{\boldsymbol{\alpha}}}$, SMI can be simply approximated as follows (Suzuki and Sugiyama, 2010):

$$\begin{aligned} \widehat{\mathrm{SMI}}(X, \widehat{E}) &= \frac{1}{n} \sum_{i=1}^n \widehat{r}(x_i, \widehat{e}_i) - \frac{1}{2n^2} \sum_{i=1}^n \widehat{r}(x_i, \widehat{e}_i)^2 - \frac{1}{2} \\ &= \widehat{\boldsymbol{h}}^\top \widehat{\boldsymbol{\alpha}} - \frac{1}{2} \widehat{\boldsymbol{\alpha}}^\top \widehat{\boldsymbol{H}} \widehat{\boldsymbol{\alpha}} - \frac{1}{2}. \end{aligned} \tag{6}$$

Input: $\{(x_i, \widehat{e}_i)\}_{i=1}^n$, $\{\sigma_i\}_{i=1}^p$, and $\{\lambda_j\}_{j=1}^q$
Output: LSMI parameter $\widehat{\boldsymbol{\alpha}}$

Compute CV score for $\{\sigma_i\}_{i=1}^p$ and $\{\lambda_j\}_{j=1}^q$ by Eq.(7);
Choose $\widehat{\sigma}$ and $\widehat{\lambda}$ that minimize the CV score;
Compute $\widehat{\boldsymbol{\alpha}}$ by Eq.(5) with $\widehat{\sigma}$ and $\widehat{\lambda}$;

Figure 1: Pseudo code of the LSMI algorithm with CV.

2.3 Model Selection in LSMI

LSMI contains three tuning parameters: the number of basis functions b, the kernel width σ, and the regularization parameter λ. In our experiments, we fix $b = \min(200, n)$, and choose σ and λ by cross-validation (CV) with grid search as follows. First, the samples $\mathcal{Z} = \{z_i \mid z_i = (x_i, \widehat{e}_i)\}_{i=1}^n$ are divided into K disjoint subsets $\{\mathcal{Z}_k\}_{k=1}^K$ of (approximately) the same size (we set $K = 2$ in experiments). Then, an estimator $\widehat{\boldsymbol{\alpha}}_{\mathcal{Z}_k}$ is obtained using $\mathcal{Z}\backslash\mathcal{Z}_k$ (i.e., without $\mathcal{Z}_k$), and the approximation error for the hold-out samples $\mathcal{Z}_k$ is computed as

$$J_{\mathcal{Z}_k}^{(K\text{-CV})} = \frac{1}{2}\widehat{\boldsymbol{\alpha}}_{\mathcal{Z}_k}^\top \widehat{\boldsymbol{H}}_{\mathcal{Z}_k} \widehat{\boldsymbol{\alpha}}_{\mathcal{Z}_k} - \widehat{\boldsymbol{h}}_{\mathcal{Z}_k}^\top \widehat{\boldsymbol{\alpha}}_{\mathcal{Z}_k},$$

where, for $|\mathcal{Z}_k|$ being the number of samples in the subset $\mathcal{Z}_k$,

$$\widehat{\boldsymbol{H}}_{\mathcal{Z}_k} = \frac{1}{|\mathcal{Z}_k|^2} \sum_{x,\widehat{e}\in\mathcal{Z}_k} \boldsymbol{\varphi}(x, \widehat{e})\boldsymbol{\varphi}(x, \widehat{e})^\top,$$

$$\widehat{\boldsymbol{h}}_{\mathcal{Z}_k} = \frac{1}{|\mathcal{Z}_k|} \sum_{(x,\widehat{e})\in\mathcal{Z}_k} \boldsymbol{\varphi}(x, \widehat{e}).$$

This procedure is repeated for $k = 1, \ldots, K$, and its average $J^{(K\text{-CV})}$ is outputted as

$$J^{(K\text{-CV})} = \frac{1}{K}\sum_{k=1}^K J_{\mathcal{Z}_k}^{(K\text{-CV})}. \tag{7}$$

We compute $J^{(K\text{-CV})}$ for all model candidates (the kernel width σ and the regularization parameter λ in the current setup), and choose the model that minimizes $J^{(K\text{-CV})}$. Note that $J^{(K\text{-CV})}$ is an almost unbiased estimator of the objective function (4), where the almost-ness comes from the fact that the number of samples is reduced in the CV procedure due to data splitting (Schölkopf and Smola, 2002).

The LSMI algorithm is summarized in Figure 1.

2.4 Least-Squares Independence Regression

Given the SMI estimator (6), our next task is to learn the parameter $\boldsymbol{\beta}$ in the regression model (1) as

$$\widehat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{\beta}} \left[\widehat{\mathrm{SMI}}(X, \widehat{E}) + \frac{\gamma}{2}\boldsymbol{\beta}^\top\boldsymbol{\beta}\right].$$

We call this method *least-squares independence regression (LSIR)*.

For regression parameter learning, we simply employ a gradient descent method:

$$\boldsymbol{\beta} \longleftarrow \boldsymbol{\beta} - \eta \left(\frac{\partial \widehat{\mathrm{SMI}}(X, \widehat{E})}{\partial \boldsymbol{\beta}} + \gamma \boldsymbol{\beta} \right), \tag{8}$$

where η is a step size which may be chosen in practice by some approximate line search method such as *Armijo's rule* (Patriksson, 1999).

The partial derivative of $\widehat{\mathrm{SMI}}(X, \widehat{E})$ with respect to $\boldsymbol{\beta}$ can be approximately expressed as

$$\frac{\partial \widehat{\mathrm{SMI}}(X, \widehat{E})}{\partial \boldsymbol{\beta}} \approx \sum_{l=1}^{b} \widehat{\alpha}_l \frac{\partial \widehat{h}_l}{\partial \boldsymbol{\beta}} - \frac{1}{2} \sum_{l,l'=1}^{b} \widehat{\alpha}_l \widehat{\alpha}_{l'} \frac{\partial \widehat{H}_{l,l'}}{\partial \boldsymbol{\beta}},$$

where

$$\begin{aligned}
\frac{\partial \widehat{h}_l}{\partial \boldsymbol{\beta}} &= \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \varphi_l(x_i, \widehat{e}_i)}{\partial \boldsymbol{\beta}}, \\
\frac{\partial \widehat{H}_{l,l'}}{\partial \boldsymbol{\beta}} &= \frac{1}{n^2} \sum_{i,j=1}^{n} \left(\frac{\partial \varphi_l(x_i, \widehat{e}_j)}{\partial \boldsymbol{\beta}} \varphi_{l'}(x_j, \widehat{e}_i) + \varphi_l(x_i, \widehat{e}_j) \frac{\partial \varphi_l(x_j, \widehat{e}_i)}{\partial \boldsymbol{\beta}} \right), \\
\frac{\partial \varphi_l(x, \widehat{e})}{\partial \boldsymbol{\beta}} &= -\frac{1}{2\sigma^2} \varphi_l(x, \widehat{e})(\widehat{e} - \widehat{v}_l) \boldsymbol{\psi}(x).
\end{aligned}$$

In the above derivation, we ignored the dependence of $\boldsymbol{\beta}$ on $\widehat{e}_i$. It is possible to exactly compute the derivative in principle, but we use this approximated expression since it is computationally efficient.

We assumed that the mean of the noise E is zero. Taking into account this, we modify the final regressor as

$$\widehat{f}(x) = f_{\widehat{\boldsymbol{\beta}}}(x) + \frac{1}{n} \sum_{i=1}^{n} \left(y_i - f_{\widehat{\boldsymbol{\beta}}}(x_i) \right).$$

2.5 Model Selection in LSIR

LSIR contains three tuning parameters—the number of basis functions m, the kernel width τ, and the regularization parameter γ. In our experiments, we fix $m = \min(200, n)$, and choose τ and γ by CV with grid search as follows. First, the samples $\mathcal{Z} = \{z_i \mid z_i = (x_i, \widehat{e}_i)\}_{i=1}^{n}$ are divided into T disjoint subsets $\{\mathcal{Z}_t\}_{t=1}^{T}$ of (approximately) the same size (we set $T = 2$ in experiments). Then, an estimator $\widehat{\boldsymbol{\beta}}_{\mathcal{Z}_t}$ is obtained using $\mathcal{Z} \backslash \mathcal{Z}_t$ (i.e., without $\mathcal{Z}_t$), and the independence criterion for the hold-out samples $\mathcal{Z}_t$ is computed as

$$\widehat{I}_{\mathcal{Z}_t}^{(T\text{-CV})} = \widehat{\boldsymbol{h}}_{\mathcal{Z}_t}^{\top} \widehat{\boldsymbol{\alpha}}_{\mathcal{Z}_t} - \frac{1}{2} \widehat{\boldsymbol{\alpha}}_{\mathcal{Z}_t}^{\top} \widehat{\boldsymbol{H}}_{\mathcal{Z}_t} \widehat{\boldsymbol{\alpha}}_{\mathcal{Z}_t} - \frac{1}{2}.$$

This procedure is repeated for $t = 1, \ldots, T$, and its average $\widehat{I}^{(T\text{-CV})}$ is computed as

$$\widehat{I}^{(T\text{-CV})} = \frac{1}{T} \sum_{t=1}^{T} \widehat{I}_{\mathcal{Z}_t}^{(T\text{-CV})}. \tag{9}$$

We compute $\widehat{I}^{(T\text{-CV})}$ for all model candidates (the kernel width τ and the regularization parameter γ in the current setup), and choose the model that minimizes $\widehat{I}^{(T\text{-CV})}$.

The LSIR algorithm is summarized in Figure 2. A MATLAB® implementation of LSIR is available from

Input: $\{(x_i, y_i)\}_{i=1}^n$, $\{\tau_i\}_{i=1}^p$, and $\{\gamma_j\}_{j=1}^q$
Output: LSIR parameter $\widehat{\boldsymbol{\beta}}$

Compute CV score for all $\{\tau_i\}_{i=1}^p$ and $\{\gamma_j\}_{j=1}^q$ by Eq.(9);
Choose $\widehat{\tau}$ and $\widehat{\gamma}$ that minimize the CV score;
Compute $\widehat{\boldsymbol{\beta}}$ by gradient descent (8) with $\widehat{\tau}$ and $\widehat{\gamma}$;

Figure 2: Pseudo code of the LSIR algorithm with CV.

'`http://sugiyama-www.cs.titech.ac.jp/~yamada/lsir.html`'.

3 Causal Direction Inference by LSIR

In the previous section, we gave a dependence minimizing regression method, LSIR, that is equipped with CV for model selection. In this section, following Hoyer *et al.* (2009), we explain how LSIR can be used for causal direction inference.

Our final goal is, given i.i.d. paired samples $\{(x_i, y_i)\}_{i=1}^n$, to determine whether X causes Y or vice versa. To this end, we test whether the causal model $Y = f_Y(X) + E_Y$ or the alternative model $X = f_X(Y) + E_X$ fits the data well, where the goodness of fit is measured by independence between inputs and residuals (i.e., estimated noise). Independence of inputs and residuals may be decided in practice by the *permutation test* (Efron and Tibshirani, 1993).

More specifically, we first run LSIR for $\{(x_i, y_i)\}_{i=1}^n$ as usual, and obtain a regression function $\widehat{f}$. This procedure also provides an SMI estimate for $\{(x_i, \widehat{e}_i) \mid \widehat{e}_i = y_i - \widehat{f}(x_i)\}_{i=1}^n$. Next, we randomly permute the pairs of input and residual $\{(x_i, \widehat{e}_i)\}_{i=1}^n$ as $\{(x_i, \widehat{e}_{\kappa(i)})\}_{i=1}^n$, where $\kappa(\cdot)$ is a randomly generated permutation function. Note that the permuted pairs of samples are independent of each other since the random permutation breaks the dependency between X and $\widehat{E}$ (if exists). Then we compute SMI estimates for the permuted data $\{(x_i, \widehat{e}_{\kappa(i)})\}_{i=1}^n$ by LSMI. This random permutation process is repeated many times (in experiments, the number of repetitions is set to 1000), and the distribution of SMI estimates under the null-hypothesis (i.e., independence) is constructed. Finally, the p-value is approximated by evaluating the relative ranking of the SMI estimate computed from the original input-residual data over the distribution of SMI estimates for randomly permuted data.

In order to decide the causal direction, we compute the p-values $p_{X\to Y}$ and $p_{X\leftarrow Y}$ for both directions $X \to Y$ (i.e., X causes Y) and $X \leftarrow Y$ (i.e., Y causes X). For a given significance level δ, we determine the causal direction as follows.

- If $p_{X\to Y} > \delta$ and $p_{X\leftarrow Y} \le \delta$, the model $X \to Y$ is chosen.
- If $p_{X\leftarrow Y} > \delta$ and $p_{X\to Y} \le \delta$, the model $X \leftarrow Y$ is selected.
- If $p_{X\to Y}, p_{X\leftarrow Y} \le \delta$, then we conclude that there is no causal relation between X and Y.
- If $p_{X\to Y}, p_{X\leftarrow Y} > \delta$, perhaps our modeling assumption is not correct.

When we have prior knowledge that there exists a causal relation between X and Y but their the causal direction is unknown, we may simply compare the values of $p_{X\to Y}$ and $p_{X\leftarrow Y}$ as follows:

- If $p_{X\to Y} > p_{X\leftarrow Y}$, we conclude that X causes Y.

- Otherwise, we conclude that Y causes X.

This simplified procedure allows us to avoid the computational expensive permutation process.

In our preliminary experiments, we empirically observed that SMI estimates obtained by LSIR tend to be affected by the basis function choice in LSIR. To mitigate this problem, we run LSIR and compute an SMI estimate 5 times by randomly changing basis functions. Then the regression function that gives the smallest SMI estimate among 5 repetitions is selected and the permutation test is performed for that regression function.

4 Existing Method: HSIC Regression

In this section, we first review the *Hilbert-Schmidt independence criterion* (HSIC) (Gretton *et al.*, 2005) and point out its potential weaknesses. Then, we review *HSIC regression* (HSICR) (Mooij *et al.*, 2009).

4.1 HSIC

The *Hilbert-Schmidt independence criterion* (HSIC) (Gretton *et al.*, 2005) is a state-of-the-art measure of statistical independence based on characteristic functions (see also Feuerverger, 1993; Kankainen, 1995). Here, we review the definition of HSIC and explain its basic properties.

Let $\mathcal{F}$ be a *reproducing kernel Hilbert space* (RKHS) with reproducing kernel $K(x,x')$ (Aronszajn, 1950), and $\mathcal{G}$ be another RKHS with reproducing kernel $L(e,e')$. Let C be a *cross-covariance operator* from $\mathcal{G}$ to $\mathcal{F}$, i.e., for all $f \in \mathcal{F}$ and $g \in \mathcal{G}$,

$$\langle f, Cg\rangle_{\mathcal{F}} = \iint \left(\left[f(x) - \int f(x)p(x)\mathrm{d}x \right] \left[g(e) - \int g(e)p(e)\mathrm{d}e \right] \right) p(x,e)\mathrm{d}x\mathrm{d}e,$$

where $\langle \cdot, \cdot \rangle_{\mathcal{F}}$ denotes the inner product in $\mathcal{F}$. Thus, C can be expressed as

$$C = \iint \left(\left[K(\cdot,x) - \int K(\cdot,x)p(x)\mathrm{d}x \right] \otimes \left[L(\cdot,e) - \int L(\cdot,e)p(e)\mathrm{d}e \right] \right) p(x,e)\mathrm{d}x\mathrm{d}e,$$

where '$\otimes$' denotes the *tensor product*, and we used the reproducing properties:

$$f(x) = \langle f, K(\cdot,x)\rangle_{\mathcal{F}} \quad \text{and} \quad g(e) = \langle g, L(\cdot,e)\rangle_{\mathcal{G}}.$$

The cross-covariance operator is a generalization of the *cross-covariance matrix* between random vectors. When $\mathcal{F}$ and $\mathcal{G}$ are *universal RKHSs* (Steinwart, 2001) defined on compact domains $\mathcal{X}$ and $\mathcal{E}$, respectively, the largest singular value of C is zero if and only if x and e are independent. Gaussian RKHSs are examples of the universal RKHS.

HSIC is defined as the the squared *Hilbert-Schmidt norm* (the sum of the squared singular values) of the cross-covariance operator C:

$$\begin{aligned} \mathrm{HSIC} := & \iiiint K(x,x')L(e,e')p(x,e)p(x',e')\mathrm{d}x\mathrm{d}e\mathrm{d}x\mathrm{d}e' \\ & + \left[\iint K(x,x')p(x)p(x')\mathrm{d}x\mathrm{d}x' \right] \left[\iint L(e,e')p(e)p(e')\mathrm{d}e\mathrm{d}e' \right] \\ & - 2 \iint \left[\int K(x,x')p(x')\mathrm{d}x' \right] \left[\int L(e,e')p(e')\mathrm{d}e' \right] p(x,e)\mathrm{d}x\mathrm{d}e. \end{aligned}$$

The above expression allows one to immediately obtain an empirical estimator—with the i.i.d. samples $\mathcal{Z} = \{(x_k, e_k)\}_{k=1}^n$ following $p(x, e)$, a consistent estimator of HSIC is given as

$$\begin{aligned}\widehat{\mathrm{HSIC}}(X, E) &:= \frac{1}{n^2}\sum_{i,i'=1}^{n} K(x_i, x_{i'})L(e_i, e_{i'}) + \frac{1}{n^4}\sum_{i,i',j,j'=1}^{n} K(x_i, x_{i'})L(e_j, e_{j'}) \\ &\quad - \frac{2}{n^3}\sum_{i,j,k=1}^{n} K(x_i, x_k)L(e_j, e_k) \\ &= \frac{1}{n^2}\mathrm{tr}(\boldsymbol{K}\boldsymbol{\Gamma}\boldsymbol{L}\boldsymbol{\Gamma}), \end{aligned} \tag{10}$$

where

$$\boldsymbol{K}_{i,i'} = K(x_i, x_{i'}), \quad \boldsymbol{L}_{j,j'} = L(e_i, e_{i'}), \quad \text{and} \quad \boldsymbol{\Gamma} = \boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top.$$

$\boldsymbol{I}_n$ denotes the n-dimensional identity matrix, and $\mathbf{1}_n$ denotes the n-dimensional vector with all ones.

$\widehat{\mathrm{HSIC}}$ depends on the choice of the universal RKHSs $\mathcal{F}$ and $\mathcal{G}$. In the original HSIC paper (Gretton *et al.*, 2005), the Gaussian RKHS with width set to the median distance between samples was used, which is a popular heuristic in the kernel method community (Schölkopf and Smola, 2002). However, to the best of our knowledge, there is no strong theoretical justification for this heuristic. On the other hand, the LSMI method is equipped with cross-validation, and thus all the tuning parameters such as the Gaussian width and the regularization parameter can be optimized in an objective and systematic way. This is an advantage of LSMI over HSIC.

4.2 HSIC Regression

In *HSIC regression* (HSICR) (Mooij *et al.*, 2009), the following linear model is employed:

$$f_{\boldsymbol{\theta}}(x) = \sum_{l=1}^{n} \theta_l \phi_l(x) = \boldsymbol{\theta}^\top \boldsymbol{\phi}(x), \tag{11}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^\top$ are regression parameters and $\boldsymbol{\phi}(x) = (\phi_1(x), \ldots, \phi_n(x))^\top$ are basis functions. Mooij *et al.* (2009) proposed to use the Gaussian basis function:

$$\phi_l(x) = \exp\left(-\frac{(x - x_l)^2}{2\rho^2}\right),$$

where the kernel width ρ is set to the median distance between points in the samples:

$$\rho = 2^{-1/2}\mathrm{median}(\{\|x_i - x_j\|\}_{i,j=1}^n).$$

Given the HSIC estimator (10), the parameter $\boldsymbol{\theta}$ in the regression model (11) is obtained by

$$\widehat{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta}} \left[\widehat{\mathrm{HSIC}}(X, Y - f_{\boldsymbol{\theta}}(X)) + \frac{\xi}{2}\boldsymbol{\theta}^\top\boldsymbol{\theta}\right], \tag{12}$$

where $\xi \geq 0$ is the regularization parameter for avoiding overfitting.

In the HSIC estimator, the Gaussian kernels,

$$K(x,x') = \exp\left(-\frac{(x-x')^2}{2\sigma_{\mathrm{x}}^2}\right) \quad \text{and} \quad L(e,e') = \exp\left(-\frac{(e-e')^2}{2\sigma_{\mathrm{e}}^2}\right),$$

are used and their kernel widths are set to the median distance between points in the samples:

$$\sigma_{\mathrm{x}} = 2^{-1/2}\mathrm{median}(\{\|x_i - x_j\|\}_{i,j=1}^n),$$
$$\sigma_{\mathrm{e}} = 2^{-1/2}\mathrm{median}(\{\|e_i - e_j\|\}_{i,j=1}^n).$$

The optimization problem (12) can be efficiently solved by using the *L-BFGS quasi-Newton method* (Liu and Nocedal, 1989) or gradient descent.

Then, the final regressor is given as

$$\widehat{f}(x) = f_{\widehat{\boldsymbol{\theta}}}(x) + \frac{1}{n}\sum_{i=1}^{n}\left(y_i - f_{\widehat{\boldsymbol{\theta}}}(x_i)\right).$$

Note that, since it is not allowed to change the kernel width σ_{e} during the optimization (12), σ_{e} is fixed to an estimate obtained based on an initial rough estimate of the residuals. This fact implies that, if the estimation accuracy of σ_{e} is poor, the overall performance of HSICR will be degraded. On the other hand, the LSIR method is equipped with cross-validation, and thus all the tuning parameters can be optimized in an objective and systematic way. This is a significant advantage of LSIR over HSICR.

5 Experiments

In this section, we first illustrate the behavior of LSIR using a toy example, and then we evaluate the performance of LSIR using benchmark datasets and real-world gene expression data.

5.1 Illustrative Examples

Let us consider the following additive noise model:

$$Y = X^3 + E,$$

where X is subject to the uniform distribution on $(-1,1)$ and E is subject to the exponential distribution with rate parameter 1 (and its mean is adjusted to have mean zero). We drew 300 paired samples of X and Y following the above generative model (see Figure 3), where the ground truth is that X and E are independent of each other. Thus, the null-hypothesis should be accepted (i.e., the p-values should be large).

Figure 3 depicts the regressor obtained by LSIR, giving a good approximation to the true function. We repeated the experiment 1000 times with the random seed changed. For the significance level 5%, LSIR successfully accepted the null-hypothesis 992 times out of 1000 runs.

As Mooij *et al.* (2009) pointed out, beyond the fact that the p-values frequently exceed the pre-specified significance level, it is important to have a wide margin beyond the significance level in order to cope with, e.g., multiple variable cases. Figure 4(a) depicts the histogram of $p_{X\to Y}$ obtained by LSIR over 1000 runs. The plot shows that LSIR tends to produce much larger p-values than the significance level; the mean and standard deviation of the p-values over 1000 runs are 0.6114 and 0.2327, respectively.

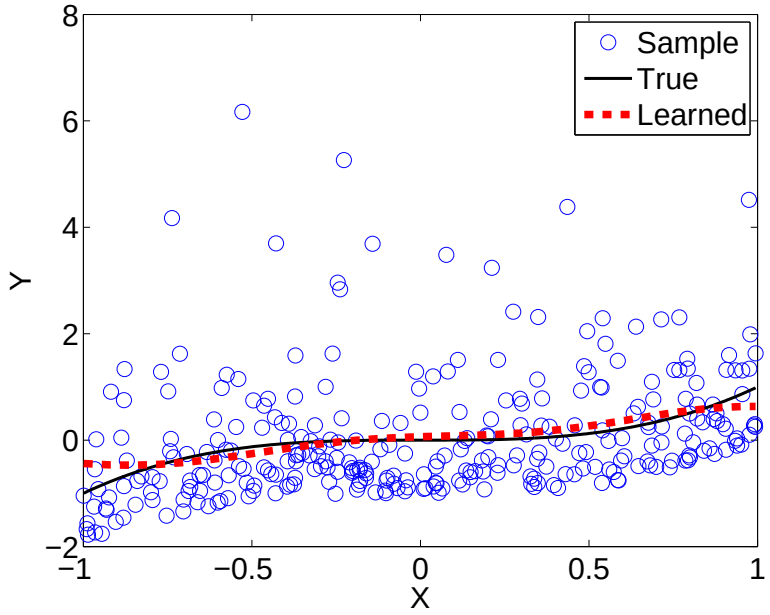


Figure 3: Illustrative example. The solid line denotes the true function, the circles denote samples, and the dashed line denotes the regressor obtained by LSIR.

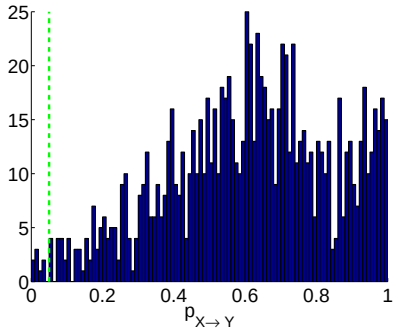
Next, we consider the backward case where the roles of X and Y were swapped. In this case, the ground truth is that the input and the residual are dependent (see Figure 3). Therefore, the null-hypothesis should be rejected (i.e., the p-values should be small). Figure 4(b) shows the histogram of $p_{X\leftarrow Y}$ obtained by LSIR over 1000 runs. LSIR rejected the null-hypothesis 989 times out of 1000 runs; the mean and standard deviation of the p-values over 1000 runs are 0.0035 and 0.0094, respectively.

Figure 4(c) depicts the p-values for both directions in a trial-wise manner. The graph shows that LSIR perfectly estimates the correct causal direction (i.e., $p_{X\to Y} > p_{X\leftarrow Y}$), and the *margin* between $p_{X\to Y}$ and $p_{X\leftarrow Y}$ seems to be clear (i.e., most of the points are clearly below the diagonal line). This illustrates the usefulness of LSIR in causal direction inference.

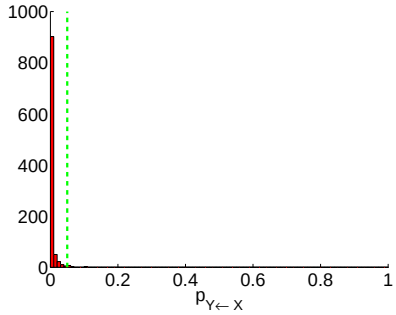
Finally, we investigate the values of independence measure $\widehat{\mathrm{SMI}}$, which are plotted in Figure 4(d) again in a trial-wise manner. The graph implies that the values of $\widehat{\mathrm{SMI}}$ may be simply used for determining the causal direction, instead of the p-values. Indeed, the correct causal direction (i.e., $\widehat{\mathrm{SMI}}_{X\to Y} < \widehat{\mathrm{SMI}}_{X\leftarrow Y}$) can be found 999 times out of 1000 trials by this simplified method. This would be a practically useful heuristic since we can avoid performing the computationally intensive permutation test.

5.2 Benchmark Datasets

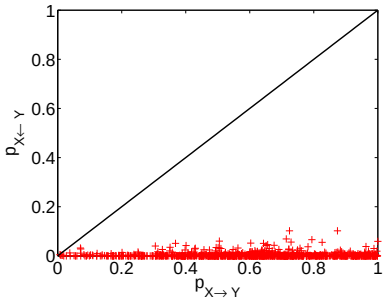
Next, we evaluate the performance of LSIR on the '*Cause-Effect Pairs*' task in the *NIPS 2008 Causality Competition* (Mooij *et al.*, 2008). The task contains 8 datasets (see Figure 5), each has two statistically dependent random variables possessing inherent causal relationship. The goal is to identify the causal direction from the observational data. Since these datasets consist of real-world samples, our modeling assumption may be only



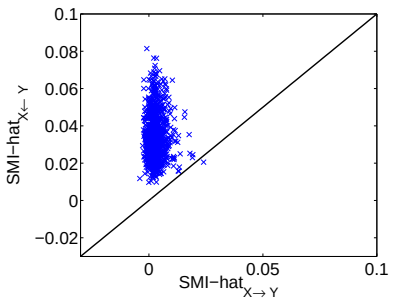
(a) Histogram of $p_{X\to Y}$ obtained by LSIR over 1000 runs. The ground truth is to accept the null-hypothesis (thus the p-values should be large).



(b) Histograms of $p_{X\leftarrow Y}$ obtained by LSIR over 1000 runs. The ground truth is to reject the null-hypothesis (thus the p-values should be small).



(c) Comparison of p-values for both directions ($p_{X\to Y}$ vs. $p_{X\leftarrow Y}$). Points below the diagonal line indicate successful trials.



(d) Comparison of values of independence measures for both directions ($\widehat{\mathrm{SMI}}_{X\to Y}$ vs. $\widehat{\mathrm{SMI}}_{X\leftarrow Y}$). Points above the diagonal line indicate successful trials.

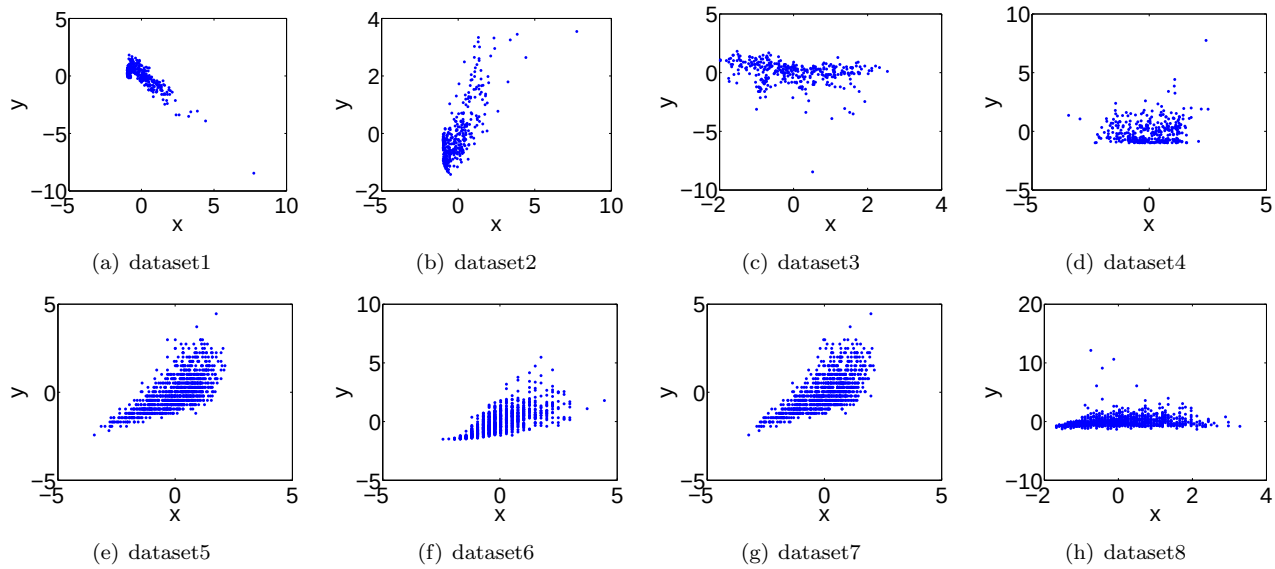
Figure 4: LSIR performance statistics in illustrative example.

approximately satisfied. Thus, identifying causal directions in these datasets would be highly challenging.

The p-values and the independence scores for each dataset and each direction are summarized in Table 1. The values of HSICR, which were also computed by the permutation test, were taken from (Mooij *et al.*, 2009), but the p-values were rounded off to three decimal places to be consistent with the results of LSIR. When the p-values of both directions are less than 10^{-3}, we concluded that the causal direction cannot be determined (indicated by '?').

Table 1 shows that LSIR successfully found the correct causal direction for 7 out of 8 cases, while HSICR gave the correct decision only for 5 out of 8 cases. This implies that LSIR compares favorably with HSICR.

The values of independence measures described in Table 1 show that merely comparing the values of $\widehat{\mathrm{SMI}}$ is again sufficient for deciding the correct causal direction in LSIR (see the estimated causal directions described in the brackets). Actually, this heuristic also allows us to correctly identify the causal direction in Dataset 8. On the other hand, in HSICR, this convenient heuristic is not as useful as in the case of LSIR.



(a) dataset1 (b) dataset2 (c) dataset3 (d) dataset4

(e) dataset5 (f) dataset6 (g) dataset7 (h) dataset8

Figure 5: Datasets of the '*Cause-Effect Pairs*' task in the *NIPS 2008 Causality Competition* (Mooij *et al.*, 2008).

5.3 Gene Function Regulations

Finally, we apply our proposed LSIR method to the real-world biological datasets, which contain known causal relationships about gene function regulations from transcription factors to gene expressions.

Causal prediction is biologically and medically important because it gives us a clue for disease-causing genes or drug-target genes. Transcription factors regulate expression levels of their relating genes. In other words, when the expression level of transcription factor genes is high, genes regulated by the transcription factor become highly expressed or suppressed.

In this experiment, we select 10 well-known gene regulation relationships of *E. coli* (Faith *et al.*, 2007), where each data contains expression levels of the genes over 445 different environments (i.e., 445 samples, see Figure 6)

The experimental results are summarized in Table 2, showing that LSIR successfully found the correct causal direction for 7 out of 10 cases, while HSICR gave the correct decision only for 4 out of 10 cases. Moreover, the causal direction can be efficiently chosen 9 out of 10 cases just by comparing the values of $\widehat{\mathrm{SMI}}$.

6 Conclusions

In this paper, we proposed a new method of dependence minimization regression called *least-squares independence regression* (LSIR). LSIR adopts the *squared-loss mutual information* as an independence measure, and it is estimated by the method of *least-squares mutual information* (LSMI). Since LSMI provides an analytic-form solution, we can explicitly compute the gradient of the LSMI estimator with respect to regression parameters. A notable advantage of the proposed LSIR method over the state-of-the-art method of dependence minimization regression (Mooij *et al.*, 2009) is that LSIR is equipped with a natural cross-validation procedure, allowing us to objectively optimize tuning parameters such as the kernel width and the regularization parameter in a data-dependent fashion. We experimentally showed that LSIR is promising in real-world causal direction inference.

Table 1: Results for the '*Cause-Effect Pairs*' task in the *NIPS 2008 Causality Competition* (Mooij *et al.*, 2008). When the p-values of both directions are less than 10^{-3}, we concluded that the causal direction cannot be determined (indicated by '?'). Estimated directions in the brackets are determined based on comparing the values of $\widehat{\mathrm{SMI}}$ or $\widehat{\mathrm{HSIC}}$.

(a) LSIR

Dataset	p-values		$\widehat{\mathrm{SMI}}$		Direction		
	$X \to Y$	$X \leftarrow Y$	$X \to Y$	$X \leftarrow Y$	Estimated		Truth
1	0.031	$< 10^{-3}$	0.0057	0.0265	$\to$	$(\to)$	$\to$
2	0.004	$< 10^{-3}$	0.0182	0.0301	$\to$	$(\to)$	$\to$
3	0.099	0.009	0.0090	0.0147	$\to$	$(\to)$	$\to$
4	0.102	0.173	0.0075	0.0051	$\leftarrow$	$(\leftarrow)$	$\leftarrow$
5	$< 10^{-3}$	0.012	0.0234	0.0108	$\leftarrow$	$(\leftarrow)$	$\leftarrow$
6	0.058	0.001	0.0079	0.0154	$\to$	$(\to)$	$\to$
7	0.009	0.018	0.0121	0.0110	$\leftarrow$	$(\leftarrow)$	$\leftarrow$
8	$< 10^{-3}$	$< 10^{-3}$	0.0149	0.0244	?	$(\to)$	$\to$

(b) HSICR

Dataset	p-values		$\widehat{\mathrm{HSIC}}$		Direction		
	$X \to Y$	$X \leftarrow Y$	$X \to Y$	$X \leftarrow Y$	Estimated		Truth
1	0.290	$< 10^{-3}$	0.0012	0.0060	$\to$	$(\to)$	$\to$
2	0.037	0.014	0.0020	0.0021	$\to$	$(\to)$	$\to$
3	0.045	0.003	0.0019	0.0026	$\to$	$(\to)$	$\to$
4	0.376	0.012	0.0011	0.0023	$\to$	$(\to)$	$\leftarrow$
5	$< 10^{-3}$	0.160	0.0028	0.0005	$\leftarrow$	$(\leftarrow)$	$\leftarrow$
6	$< 10^{-3}$	$< 10^{-3}$	0.0032	0.0026	?	$(\leftarrow)$	$\to$
7	$< 10^{-3}$	0.272	0.0021	0.0005	$\leftarrow$	$(\leftarrow)$	$\leftarrow$
8	$< 10^{-3}$	$< 10^{-3}$	0.0015	0.0017	?	$(\to)$	$\to$

Acknowledgments

MY was supported by the JST PRESTO program and MS was supported by SCAT, AOARD, and the FIRST program.

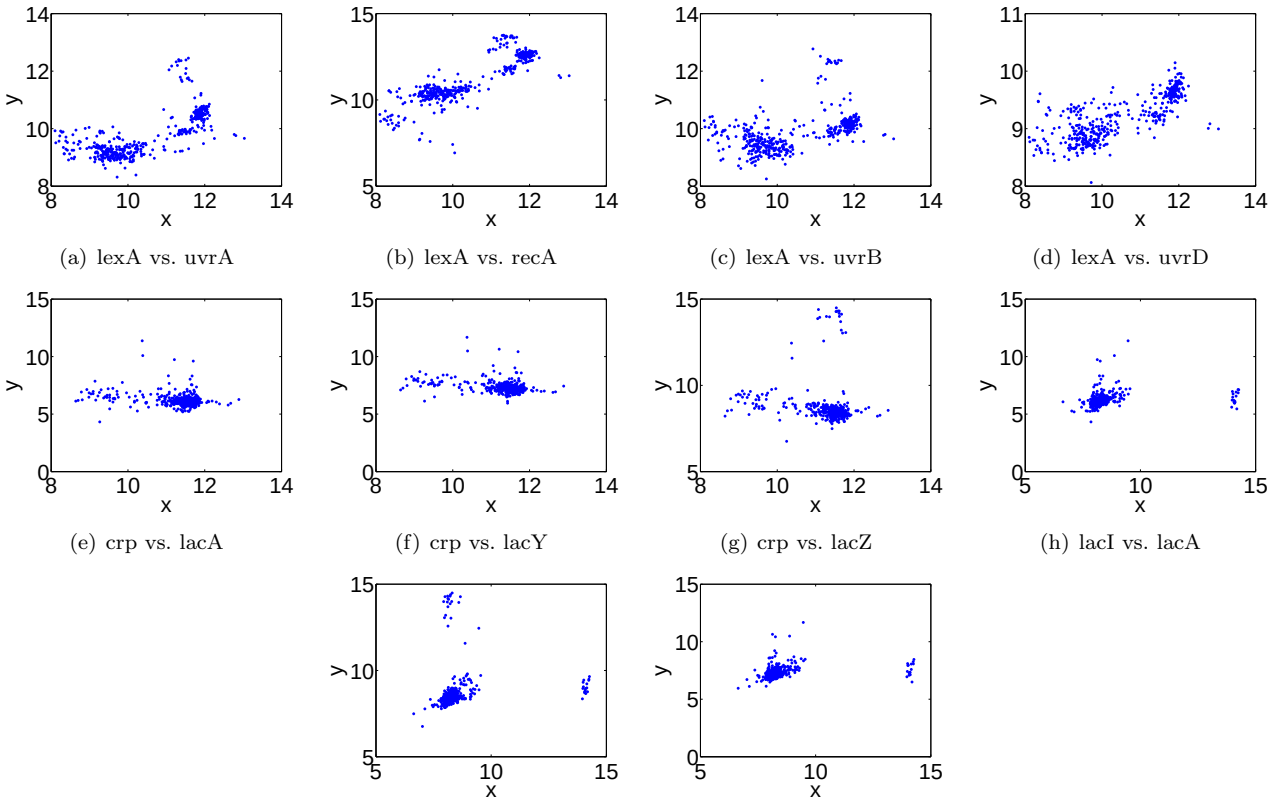
References

Aronszajn, N. (1950). Theory of reproducing kernels. *Trans. the American Mathematical Society*, **68**, 337–404.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York, NY.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. John Wiley & Sons, Inc., Hoboken, NJ, USA, 2nd edition.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York, NY.



(a) lexA vs. uvrA (b) lexA vs. recA (c) lexA vs. uvrB (d) lexA vs. uvrD

(e) crp vs. lacA (f) crp vs. lacY (g) crp vs. lacZ (h) lacI vs. lacA

(i) lacI vs. lacZ (j) lacI vs. lacY

Figure 6: Datasets of the E. coli task (Faith *et al.*, 2007).

Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif1, S., Collins, J. J., and Gardner, T. S. (2007). Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biology*, **5**(1), e8.

Feuerverger, A. (1993). A consistent test for bivariate dependence. *International Statistical Review*, **61**(3), 419–433.

Fukumizu, K., Bach, F. R., and Jordan, M. (2009). Kernel dimension reduction in regression. *The Annals of Statistics*, **37**(4), 1871–1905.

Geiger, D. and Heckerman, D. (1994). Learning Gaussian networks. In *10th Annual Conference on Uncertainty in Artificial Intelligence (UAI1994)*, pages 235–243.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005). Measuring statistical dependence with Hilbert-Schmidt norms. In *16th International Conference on Algorithmic Learning Theory (ALT 2005)*, pages 63–78.

Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. (2009). Nonlinear causal discovery with

Table 2: Results for the 'E. coli' task. When the p-values of both directions are less than 10^{-3}, we concluded that the causal direction cannot be determined (indicated by '?'). Estimated directions in the brackets are determined based on comparing the values of $\widehat{\mathrm{SMI}}$ or $\widehat{\mathrm{HSIC}}$.

(a) LSIR

Dataset		p-values		$\widehat{\mathrm{SMI}}$		Direction		
X	Y	$X \to Y$	$X \leftarrow Y$	$X \to Y$	$X \leftarrow Y$	Estimated		Truth
lexA	uvrA	$< 10^{-3}$	$< 10^{-3}$	0.0177	0.0255	?	$(\to)$	$\to$
lexA	recA	0.024	0.061	0.0070	0.0053	$\leftarrow$	$(\leftarrow)$	$\to$
lexA	uvrB	$< 10^{-3}$	$< 10^{-3}$	0.0172	0.0356	?	$(\to)$	$\to$
lexA	uvrD	0.043	$< 10^{-3}$	0.0075	0.0227	$\to$	$(\to)$	$\to$
crp	lacA	0.143	$< 10^{-3}$	-0.0004	0.0399	$\to$	$(\to)$	$\to$
crp	lacY	0.003	$< 10^{-3}$	0.0118	0.0247	$\to$	$(\to)$	$\to$
crp	lacZ	0.001	$< 10^{-3}$	0.0122	0.0307	$\to$	$(\to)$	$\to$
lacI	lacA	0.787	$< 10^{-3}$	-0.0076	0.0184	$\to$	$(\to)$	$\to$
lacI	lacZ	0.002	$< 10^{-3}$	0.0096	0.0141	$\to$	$(\to)$	$\to$
lacI	lacY	0.746	$< 10^{-3}$	-0.0082	0.0217	$\to$	$(\to)$	$\to$

(b) HSICR

Dataset		p-values		$\widehat{\mathrm{HSIC}}$		Direction		
X	Y	$X \to Y$	$X \leftarrow Y$	$X \to Y$	$X \leftarrow Y$	Estimated		Truth
lexA	uvrA	$< 10^{-3}$	$< 10^{-3}$	0.0865	0.1990	?	$(\to)$	$\to$
lexA	recA	$< 10^{-3}$	$< 10^{-3}$	0.2129	0.1625	?	$(\leftarrow)$	$\to$
lexA	uvrB	0.005	$< 10^{-3}$	0.0446	0.1335	$\to$	$(\to)$	$\to$
lexA	uvrD	$< 10^{-3}$	$< 10^{-3}$	0.0856	0.2427	?	$(\to)$	$\to$
crp	lacA	0.006	$< 10^{-3}$	0.0362	0.1162	$\to$	$(\to)$	$\to$
crp	lacY	$< 10^{-3}$	$< 10^{-3}$	0.0393	0.1303	?	$(\to)$	$\to$
crp	lacZ	$< 10^{-3}$	$< 10^{-3}$	0.0832	0.0836	?	$(\to)$	$\to$
lacI	lacA	0.004	$< 10^{-3}$	0.0368	0.1076	$\to$	$(\to)$	$\to$
lacI	lacZ	$< 10^{-3}$	$< 10^{-3}$	0.0666	0.1365	?	$(\to)$	$\to$
lacI	lacY	0.026	$< 10^{-3}$	0.0303	0.0927	$\to$	$(\to)$	$\to$

additive noise models. In D. Koller, D. Schuurmans, Y. Bengio, and L. Botton, editors, *Advances in Neural Information Processing Systems 21 (NIPS2008)*, pages 689–696, Cambridge, MA. MIT Press.

Kanamori, T., Suzuki, T., and Sugiyama, M. (2009). Condition number analysis of kernel-based density ratio estimation. Technical report, arXiv. http://www.citebase.org/abstract?id=oai:arXiv.org:0912.2800.

Kankainen, A. (1995). *Consistent Testing of Total Independence Based on the Empirical Characteristic Function*. Ph.D. thesis, University of Jyväskylä, Jyväskylä, Finland.

Kraskov, A., Stögbauer, H., and Grassberger, P. (2004). Estimating mutual information. *Physical Review E*, **69**(066138).

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, **22**, 79–86.

Liu, D. C. and Nocedal, J. (1989). On the limited memory method for large scale optimization. *Mathematical Programming B*, **45**, 503–528.

Mooij, J., Janzing, D., and Schölkopf, B. (2008). Distinguishing between cause and effect. http://www.kyb.tuebingen.mpg.de/bs/people/jorism/causality-data/.

Mooij, J., Janzing, D., Peters, J., and Schölkopf, B. (2009). Regression by dependence minimization and its application to causal inference in additive noise models. In *26th Annual International Conference on Machine Learning (ICML2009)*, pages 745–752, Montreal, Canada.

Patriksson, M. (1999). *Nonlinear Programming and Variational Inequality Problems*. Kluwer Academic, Dredrecht.

Pearl, J. (2000). *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, NY, USA.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, **50**, 157–175.

Schölkopf, B. and Smola, A. J. (2002). *Learning with Kernels*. MIT Press, Cambridge, MA.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. J. (2006). A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, **7**, 2003–2030.

Steinwart, I. (2001). On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, **2**, 67–93.

Suzuki, T. and Sugiyama, M. (2010). Sufficient dimension reduction via squared-loss mutual information estimation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS2010)*, volume 9 of *JMLR Workshop and Conference Proceedings*, pages 804–811.

Suzuki, T., Sugiyama, M., Kanamori, T., and Sese, J. (2009). Mutual information estimation reveals global associations between stimuli and biological processes. *BMC Bioinformatics*, **10**(S52).

Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley, New York, NY.